Data formats in libraries

LIBR 509
Week 9 Mar 15 2022

1 Recorded Lecture

Introduction to data formats (what are they? What effect do they have on the rest of the system's operation?), and a closer look at MARC

1.1 Introduction to Data Formats

We previously talked about **Dublin Core** as a pretty common content schema (reminder: a schema tells us what is worth describing about the object, offers some labels of how to identify what you're saying about a resource)

- Dublin Core could be written in a CSV file (column A=Attribute, column B=Value)

Attribute	Value
Title	Untitled Goose Game
Creator	House House
Subject	Geese; Mischief
Description	It's a lovely morning in the village, and you are a horrible goose.
Publisher	Panic, Inc.
Date	September 20, 2019

- Or could be written in XML data format (same information as the CSV, but different data format)

```
<dc:title>Untitled Goose Game</dc:title>
<dc.creator>House House</dc.creator>
<dc.subject>Geese</dc.subject>
<dc.subject>Mischief</dc.subject>
<dc:description> It is a lovely day in the village, and you are a horrible
goose. </dc:description>
<dc:publisher>Panic, Inc.</dc:publisher>
<dc.date>September 20, 2019</dc.date>
```

- MARC (most common data format used in contemporary libraries)

```
110  2 |a House House
245 10 |a Untitled Goose Game c| by House House
260    |a Melbourne, Australia |b Panic, Inc. c| 2019
520    |a It's a lovely morning in the village, and you are a horrible goose.
       c| Item packaging
650  4 |a Geese
650  4 |a Mischief
```

- Or on an index card (notice no element labels—determining what each piece of information represents is entirely interpreted by norms about formatting and spacing)

1.2 Machine Readable Cataloguing for the 21st Century (MARC21)

MARC records look something like this (series of 3-digit codes, spaces or additional numbers, and series of pieces of data separated by letter and number codes):

Technological uncertainty and the pure theory of allocation : an essay / N.F. Laing.

000 01925cam a2200421 i 4500
001 238606
005 20171003101234.0
008 961220s1978 xx a b 001 0 eng
010 __ |**a** 79307729
020 __ |**a** 0959589406
020 __ |**a** 9780959589405
035 __ |**a** (OCoLC)ocm05288648
035 __ |**z** U00002683316 |**i** 30oc96nw[r]
035 __ |**9** AAX-9852
035 __ |**a** (OCoLC)5288648
040 __ |**a** DLC |**b** eng |**c** DLC |**d** AUT |**d** OCLCQ |**d** LVB |**d** OCLCQ |**d** OCLCF |**d** OCLCQ |**d** OCLCO |**d** MBB |**d** OCLCQ
050 04 |**a** HC79.T4 |**b** L34
082 0_ |**a** 338/.06
090 __ |**a** HC79.T4 |**b** L34 1978
093 99 |**a** HC79.T7 |**b** L34 1978
100 1_ |**a** Laing, N. F.
245 10 |**a** Technological uncertainty and the pure theory of allocation : |**b** an essay / |**c** N.F. Laing.
246 3_ |**a** Uncertainty and allocation
260 __ |**a** [Place of publication not identified] : |**b** [publisher not identified], |**c** 1978 |**e** ([Bedford Park, Australia] : |**f** University Relations Unit, Flinders University of South Australia)
300 __ |**a** 151 pages : |**b** illustrations ; |**c** 24 cm
336 __ |**a** text |**b** txt |**2** rdacontent
337 __ |**a** unmediated |**b** n |**2** rdamedia

- Note: MARC is the exception to thinking about data formats and content standards as separate; more than any other data format, MARC already anticipates a certain kind of data that can go in each space; if you make a change to a content standard, MARC itself has to change to make a place for that information

1.2.1 Indicators and Subfields

An **Indicator** tells the computer what kind of information is coming

Subfields break down that information into its component parts, so they can be interpreted and recalled in different ways by the software

Example MARC record 245 field:

245 10
|**a** Melancholy baby : |**b** the unplanned consequences of the G.I
Winfield.
260 __ |**a** Westport, Conn. : |**b** Bergin & Garvey, |**c** 2000.
300 __ |**a** xv, 158 p. ; |**c** 24 cm.

- “245”: what kind of information this row contains (title)

- “10”: indicator: tells computer something about what kind of information is coming
- “a”, “b”: subfields: break down information into its components (a=title, b=subtitle, c=statement of responsibility)

Example: (a previous textbook for this course)

```
100  1_ |a Taylor, Arlene G.,  |d 1941-
245  14 |a The organization of information /  |c Arlene G. Taylor and Daniel N. Joudrey.
250  __ |a 3rd ed.
260  __ |a Westport, Conn. :  |b Libraries Unlimited,  |c 2009.
300  __ |a xxvi, 512 p. :  |b ill. ;  |c 26 cm.
490  1_ |a Library and information science text series
504  __ |a Includes bibliographical references (p. 479-498) and index.
650  _0 |a Information organization.
650  _0 |a Metadata.
700  1_ |a Joudrey, Daniel N.
830  _0 |a Library and information science text series.
```

245: indicates a title
1: telling the computer something about the placement of the title in the record as a whole
4: telling computer how to alphabetize (“skip the first 4 characters”)
a: title
c: statement of responsibility

Some other fields you will see reliably:

- 100, 110, or 111: the “primary creator” of a work
  - `100 (Personal Name):` item attributed to a person as the primary creator (e.g., Aristotle, Edward Short, etc.)
  - `110 (Corporation):` e.g., American Library Association, United Nations
  - `111 (Meeting):` e.g., 2022 ALA Meeting
- 700, 710, 711: additional people responsible for a work (e.g., additional authors, illustrator)
  - `700 (Personal Name)`
  - `710 (Corporation)`
  - `711 (Meeting)`

An **Indicator** tells the computer what kind of information is coming

- For 100 & 700 personal names, the first indicator tells you what type of personal name entry element (0-Forename, 1-Surname, 3-Family name)

Subfields break down that information into its component parts, so they can be interpreted and recalled in different ways by the software

- |a – Personal name (NR=”not repeatable”)

  `|b – Numeration` (e.g., “III” – the third)

  `|c – titles and other words associated with a name` (e.g., Duchess, Dr.)

`|d – Dates associated with a name` (generally year of birth) (NR); used to disambiguate people who otherwise look identical in the system, e.g. they have the same first and last name)

`|q – fuller form of name (NR)` (e.g., a fuller legal name than their personal name)

Field 245 Title Statement also has indicators and subfields:

- `Indicator 1: Title added entry`
    - `0 – No title added entry` (the title is not an additional entry for the work; the title is the first entry for the work in the database—e.g., there is no primary creator)
    - `1 – Title added entry` (the title is an additional entry in addition to the author—there is something in the 1XX field)
- Indicator 2: "Nonfiling characters" (how many characters to skip when alphabetizing)
- Subfields used most often:
    - `|a – Title (NR)`
    - `|b – Remainder of title (NR)` (e.g., subtitles)
    - `|c – statement of responsibility, etc. (NR)` (person's name, "from the mind of xx,"—transcribing the title page
        - Note: if you search an exact phrase that includes the title and the subtitle in most library databases, you won't get an exact result—this is because they are broken into different fields

`Field 650: Subject Added Entry-Topical Term (R)`

- Everything in MARC is about "aboutness", geographical coverage, chronological coverage, etc.
- Indicator 2: Thesaurus (which thesaurus it comes from)
    - `0 – Library of Congress Subject Headings`
    - `1 – LC subject headings for children's literature`
    - …
    - `7 – Source specified in subfield`
- Control subfield seen most often:
    - `|2 – Source of heading or term` (if you have a term in this list that doesn't come from any 6 that get a proper name, you can add your own)

Documentation:

- MARC 100 field: https://www.loc.gov/marc/bibliographic/bd100.html
- MARC 245 field: https://www.loc.gov/marc/bibliographic/bd245.html

Summary
Content schemas and data formats are separate things that interact with each other often

A **content schema** has:

- A set of values (attributes that are worth describing about a resource)
- Some instructions on which elements are necessary (e.g., “required fields”)
- Some instruction on how to modify elements (what to do to distinguish date published vs. date acquired vs. date digitized)
- Some instructions on how to fill out the values (e.g., what to do in the case of misspellings, inferred vs. transcribed data, etc.)

A **data format** determines:

- How to express connections between attributes and values (e.g., in CSV, they occur in the same row)
- How to express connections between attributes (e.g., contributor’s name and a contributor’s role)
- What characters you can use, how many characters you can use in a given field
- How records can relate to, overlap, add to each other
- How you can use (sort, search, filter, combine) the records
- If you are in danger of papercuts

2 Guest Lecture, Bri Watson

Linked data is the process of

E.g.,	Virginia Woolf	Wrote	A Room of One’s Own
	Subject	Predicate	Object

Example: Google is pulling data from many different sources

Libraries are now doing this too
E.g., mapping all paintings of Wales to their location; can be used by climate scientists to understand climate change

Wikidata

1 item = 1 page

Items have properties and values

A **claim** is a wrapper that includes property, value, qualifier

A **statement** is a wrapper that includes: claim and reference

3 Readings

3.1 DoO sections 9.1-9.4

This chapter is looking at the **structures of resource descriptions:** how do descriptions enable or inhibit particular ways of interacting with those descriptions?

3.1.1 Kinds of structures: blobs, sets, lists, dictionaries, trees, graphs

Blob: unstructured, no clearly defined internal parts

- Easy to create these kinds of descriptions (not necessary to think about internal parts)

Set: a collection of items

- Easy to create
- Facilitate easily finding intersections among descriptions
- Possible to introduce constraints to a set (e.g., set can only contain a maximum number of items; set has to contain a particular number of items; set items need to come from a controlled vocabulary)
- E.g., sets of tags used to describe files

List: an ordered collection of items

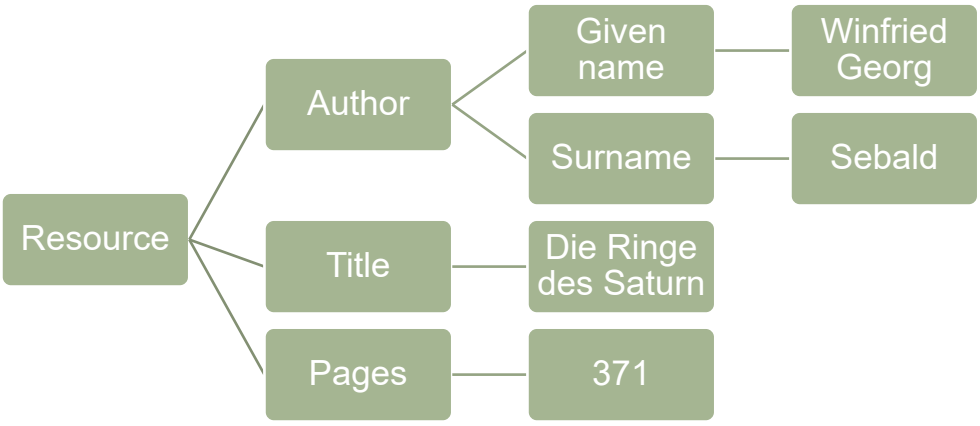
- Some possible constraints that can be applied: fixed-length; list needs to contain lists that themselves don't contain lists (a **table**)

Dictionary: a collection of key-value pairs (also called **map** or **associative array**); each entry needs to have a unique key

Tree: a nested structure consisting of **nodes** joined by **edges** (parent nodes are joined to their children by edges; nodes with no parents are **root nodes**, children with no children are **leaf nodes**)

- Some foundational properties: every node except the parent has exactly 1 parent; edges are directional (one node is a parent and one is a child)
- Some possible constraints to introduce: e.g., XML Information Sets have tree structure where children are ordered (as opposed to nested dictionaries, where children have no order)
- Good at describing different but related sources

Resource Description as Tree

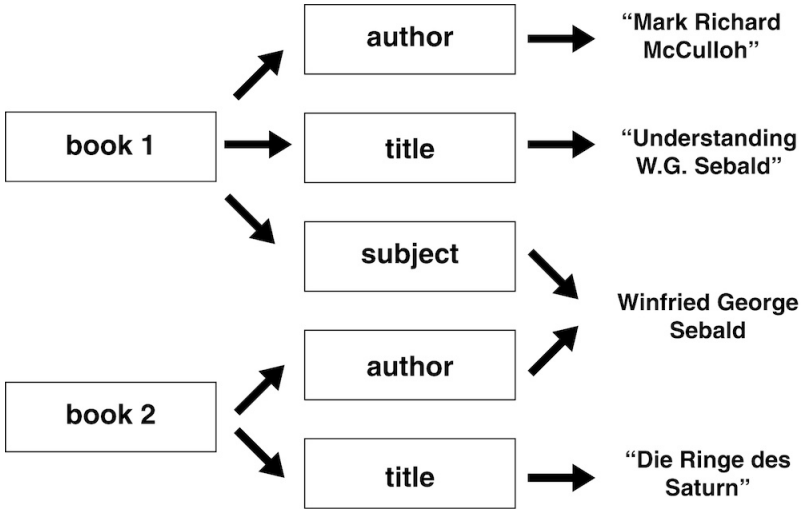


(book is primary resource; but could have also been structured where author is primary resource, with “books authored” listed underneath)

Graph: a set of nodes connected by edges; may or may not have a direction (like trees but not necessarily directed, nodes can have more than one parent)

- Very flexible structures

Resource Descriptions linked into a Graph



3.1.2 Comparing metamodels: JSON, XML and RDF

Metamodels: describe structures commonly found in resource descriptions and other information resources, regardless of the specific domain.

3.1.2.1 JSON

JSON = JavaScript Object Notation

Subset of the JavaScript programming language

Consists of 2 kinds of structures (arrays [lists] and objects [dictionaries]) that can contain various types of values (strings, numbers, Booleans, null, arrays, objects)

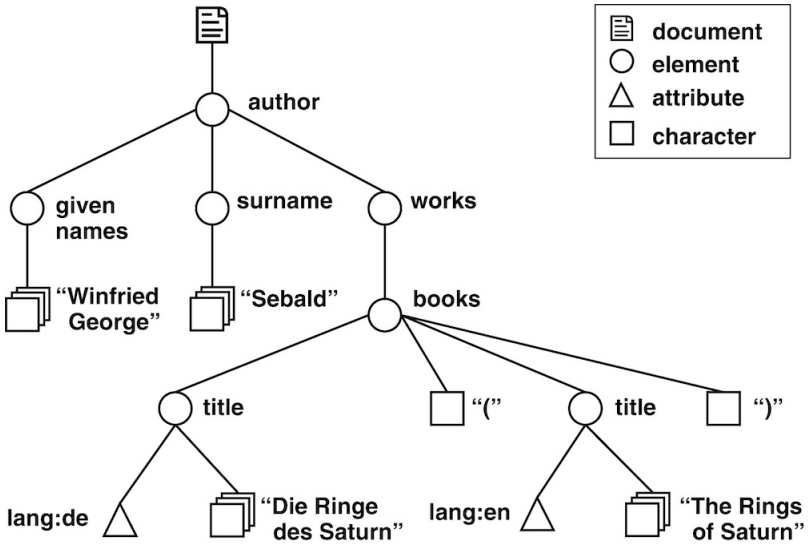
JSON is easy to work with in JavaScript (therefore popular for web applications)

3.1.2.2 XML Information Set

Derived from data structures used for document markup (**elements** and **attributes**)

XML Infoset is a tree structure: each node is an "information item" . Information item could be:

- **Document item** (root node; has exactly one element item as its child)
- **Element item**: has a set of **attribute items** and a list of child nodes (child nodes could be other element items, or character items)
- **Attribute item:** could contain "character items" or typed data



- Supports mixed content (character items and element items can be siblings)

3.1.2.3 Resource Description Framework (RDF)

RDF metamodel is a **directed graph:**

- one node is the **subject** of a **triple**, connected to its **object** by a **predicate**
- Nodes are associated with URIs as identifiers

3.1.3 Writing Descriptions

Encoding scheme: specify how to textually represent information

Writing system: uses notations, and adds a set of rules for using them (different writing systems may use the same notation differently)

Syntax: the rules that define how characters can be combined into words, and how words can be combined into higher-level structures

In the web, documents, data, and services are resources identified using URIs (Uniform Resource Identifiers) and accessible through representations transferred via the Hypertext Transfer Protocol (HTTP)

3.2 Suggested: Dana Svitavsky, An Introduction to Bibframe

https://www.youtube.com/watch?v=tFq50SW5bPk

3.3 Suggested: Kiryakos, S. and Sugimoto, S. (2019), "Building a bibliographic hierarchy for manga through the aggregation of institutional and hobbyist descriptions", Journal of Documentation, Vol. 75 No. 2, pp. 287-313.